Prediction of House Price Using Linear Regression Model

Aritra Dutta(19211293)
MSc. In Computing
Dublin City University
Dublin, Ireland
Aritra.dutta2@mail.dcu.ie

Souradip Goswami (19210273)
MSc. In Computing
Dublin City University
Dublin, Ireland
Souradip.goswami2@mail.dcu.ie

Abstract

Accurate house price prediction is of prime importance for various real estate stakeholders like house-owners, buyers, investors and agents. In this paper, we are using the King County house prices data to generate a best fit linear regression model and figure out the accuracy of the model while predicting the prices for test data. The location which plays a pivotal role in determining the price of a particular house has also been taken into account along with the in-house features. We have extracted the city from the latitude and longitude and tried to figure out how the prices correlate with a particular city.

Keywords

House price, Linear Regression model

1. Introduction

House prices play a significant role in understanding the economy and the corresponding living standard of a particular place. In Ireland, there has been a steady increase in house sales over the last decade. In USA, the house sales have increased drastically over the course of last 5 years. Thus, house price predictions have become utterly important for the real estate stakeholders for making informed decisions.

House prices are considered to be related to various features. In our study, we are trying to figure out the optimum number of features that are required to predict the house prices of King County with high accuracy based on the linear regression model. Along with the in-house features like total living area, number of bedrooms, bathrooms etc., we have also taken into account the city where the house is present, overall condition of the house and other valuable features that contribute significantly.

2. Related Work

House price can be considered as a quantitative representation of a set of features. Over the past decade, a large amount of studies has examined the relationship between the price of the house and the different features associated with it. The work of [1], discussed the importance of the location of a house in price determination.

House features can be broadly classified into two main sub-divisions: non-geographical features like area, bedroom, bathroom and geographical features like location, distance from city center and others.

The hedonic price models [2], [3] studies extensively about different features of a house. The work of [4] studies vividly about different features of houses in Nairobi and corresponding variation in the prices. Similar works have been done by Kryvobokov and Wilhelmsson [5] and predicted the market value of apartments in Donetsk, Ukraine.

The house prediction model is a hugely studied field and we are aiming to check how a simple linear regression model performs and what are the key drawbacks of such a model that prompted different researchers to look into other suitable alternative.

3. Data Profiling

In this section, we will elaborately discuss about the data that is being used for

experiment, different attributes it possesses and some cleaning and processing that were done on the data.

3.1 Data description

We have used the King County House data from Kaggle [6] as an example to predict the house prices using our model. The dataset consists of 21k rows and 25 columns. We have incorporated an additional column named 'City' which have been extracted from Google API using the latitude and longitude. The other significant columns include ID, Price, Number of bedrooms, Number of bathrooms, living area in sqft., Lot area in sqft., Number of floors, Waterfront (0= no visible waterbody, 1= water body facing), View (how many times a particular property has been viewed), Grade (3 signifying lowest and 13 signifying highest), Condition (1 = poor and 5 = Excellent), Year_built, Year_Renovated, Latitude and Longitude.

3.2 Exploratory Data Analysis

The first task before preparing the model or coming to any conclusion is to analyze the dataset, understand the various aspects of the response variable as well as determining the various correlation between the variables.

The summary of the response variable 'Price' shows that the Median is less than the Mean of the distribution. From the boxplot and the histogram density plot of price distribution, it has been observed that there are quite a few extreme values and the median is towards the 1st quantile. It is quite evident that the data is positively skewed with "visible" outliers.

The variation of the prices has been checked with respect to all the categorical predictor variables.

a) Bedrooms- The summary of the bedrooms clearly depicts that houses with 2,3 and 4 bedrooms are mostly in demand. The price increases almost linearly with increase in number of beds with few exceptions.
b) Bathrooms- The summary of the bathroom shows that most of the houses have baths in the range from 1 to 3.5. There is mostly an increase in house price with the increase of number of bathrooms which is self explanatory as more bathrooms signify increase in house size. The boxplot for each bathroom and corresponding prices shows that some outliers do exist.
c) Waterfront- The houses with waterfront has quite higher prices with respect to the ones that are not as it is clear from the summary and the boxplot between Waterfront and Price distribution.
d) Condition- The houses with better condition has higher prices. Most of the houses have moderate condition i.e. Condition=3.
e) Grade- There is a steady increase of the price with increase in Grade of the house with most of the houses having average grade points.
f) City- The prices vary quite distinctly with respect to the city.

The price distribution has been analyzed with respect to numeric predictor variables.

a) Living area- The summary of the living area shows that the distribution is positively skewed. The correlation between price and living area of the house is found to be 0.7 which signifies a strong linear relation between the response and predictor variable.
b) Lot area- The summary of the lot area shows that the mean of the distribution is greater than the median i.e. the data is positively skewed. The correlation between price and lot is found to be 0.17 which signifies a weak linear relationship.
c) Year built- The summary of the year-built shows that most of the houses were built between 1969-2015. The Pearson correlation test shows that there is almost no linear relationship between the price of the house w.r.t the year in which it was built, and the correlation coefficient is obtained as 0.036.

4. Experiment and Evaluation

In this paper we are trying to build a best fit linear regression model for predicting the

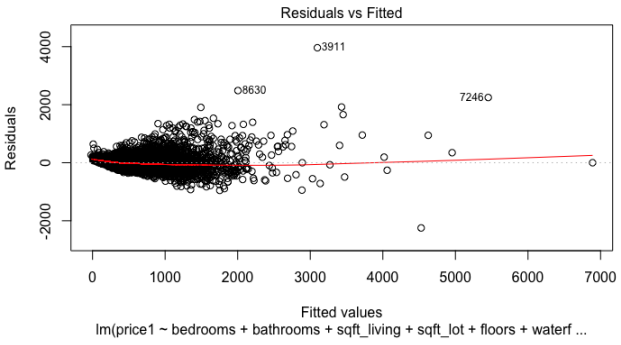
house prices in King County and discuss the performances and corresponding drawbacks.

4.1 Linear Regression Model

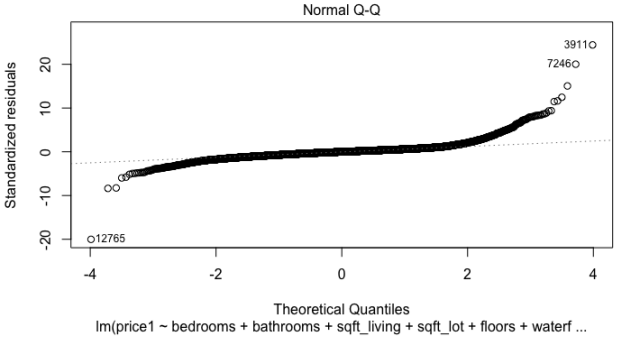
The linear model equation for the given dataset has been created using the below variables:

- Numeric: Living area, Lot area and built year
- Categorical: Bedrooms, bathrooms, floors, view, grade, waterfront, city and condition.

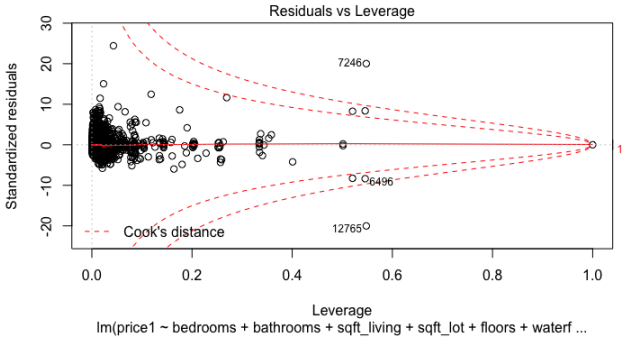
The summary of the linear model shows that the median is quite close to zero i.e. near to the mean. However, Residual vs Fitted plot (below) clearly shows that there are quite a lot of variation between the actual and predicted values although the majority of the residuals are evenly distributed along the mean with almost constant variance.



The Q-Q plot (below) suggests that the Residual values follow the 45-degree line closely but deviates by a large margin in the upper tail. This shows there are definite outliers.



The Residual vs Leverage point depicts the Cook's distance. If the value of a particular residual point is more than 0.5 then that point is regarded as influential and if the value exceeds 1, then that point is highly influential. Here we can see some influential points.



4.2 ANOVA

The ANOVA Type 2 has been performed to check whether a particular variable contribute significantly to the linear model or not.

```
Anova Table (Type II tests)

Response: price1
               Sum Sq    Df  F value    Pr(>F)
bedrooms      3648985    10   12.870 < 2.2e-16 ***
bathrooms    26813700    27   35.028 < 2.2e-16 ***
sqft_living  44876257     1 1582.848 < 2.2e-16 ***
sqft_lot       377484     1   13.314 0.0002643 ***
floors        3168892     5   22.354 < 2.2e-16 ***
waterfront   27274287     1  962.002 < 2.2e-16 ***
view         10512053     4   92.694 < 2.2e-16 ***
grade        75429762    10  266.051 < 2.2e-16 ***
condition     4103298     4   36.182 < 2.2e-16 ***
City        182441287    48  134.062 < 2.2e-16 ***
yr_built     12132712     1  427.938 < 2.2e-16 ***
Residuals   422041772 14886
```

The p-values for all the predictor variables are significantly less than .05 (null hypothesis rejected). We can conclude that all the predictor variables contribute significantly to the model and are important in determining the price of the house.

4.3 Durbin-Watson Test

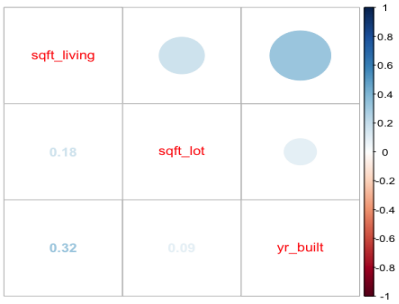
The Durbin-Watson test is performed on the linear model. This test the hypothesis that there is no autocorrelation between the samples and in turn the residuals.

```
lag Autocorrelation D-W Statistic p-value
  1    -0.003093556      2.006085   0.688
Alternative hypothesis: rho != 0
```

The above result shows that the D-W statistic is very close to 2 signifying that there is no existence of auto correlation. The p-value is 0.688 which is significantly higher than 0.05 i.e. 95 percent confidence interval.

4.4 Multi-collinearity check

Multi-collinearity leads to unstable estimation of the parameters. So, the same has been checked between different predictor variables and the result is depicted below.

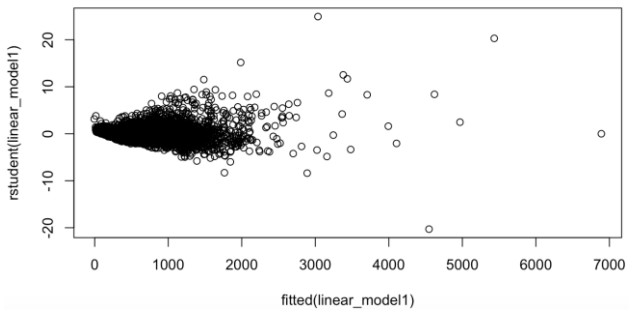


	GVIF	Df	GVIF^(1/(2*Df))
bedrooms	5.067532	10	1.084526
bathrooms	29.132050	27	1.064432
sqft_living	4.942071	1	2.223077
sqft_lot	1.223403	1	1.106076
floors	2.943775	5	1.114014
waterfront	1.597122	1	1.263773
view	1.908926	4	1.084173
grade	10.501167	10	1.124765
condition	1.386418	4	1.041686
City	4.680850	48	1.016208
yr_built	2.949151	1	1.717309

The correlation matrix which have values close to zero and the variation inflation factor (VIF) having values less than 5 shows that there is no multi collinearity between the predictor variables.

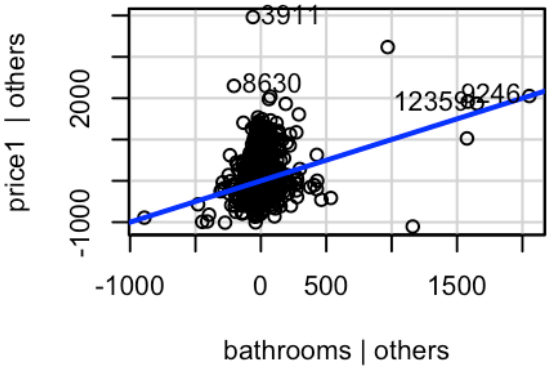
4.5 Zero Conditional Mean check

In the studentized residual plot, most of the residuals are plotted with constant variance around zero (mean). The extreme values show that some prediction have been done inaccurately.

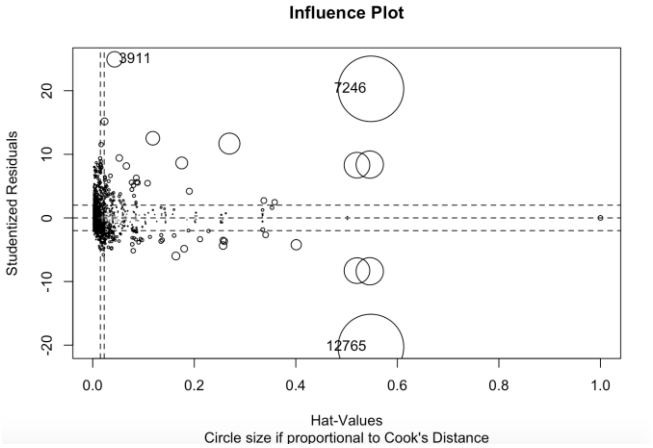


4.6 Leverage, Influence and Outliers

Leverage Point: A point in the dataset having unusual X value. The leverage plot shows that there are quite a few leverage points having unusual X values. A sample leverage plot with respect to bathroom vs price has been shown.



Influence point: An influential point is a point on the dataset, whose removal causes a large change in the fit of the model. An influential point is either an outlier or a leverage point. There are three extreme influential points in our data as suggestive from the below plot.



Outliers: They are the points which does not correspond to fitted model and leads to bad estimation. The outlier test in our model shows that there exist quite a few outliers that needs to be removed to make the model perform better.

	rstudent	unadjusted p-value	Bonferroni p
3911	24.933050	1.8409e-134	2.7593e-130
7246	20.295498	2.3448e-90	3.5146e-86
12765	-20.295498	2.3448e-90	3.5146e-86
8630	15.165516	1.4525e-51	2.1772e-47
1447	12.524860	8.2721e-36	1.2399e-31
1163	11.684373	2.1008e-31	3.1489e-27
10436	11.524863	1.3332e-30	1.9984e-26
14018	9.419979	5.1623e-21	7.7377e-17
11524	9.354140	9.6080e-21	1.4401e-16
11964	8.892038	6.6792e-19	1.0011e-14

The above table shows that there are some outliers. The Bonferroni values for the all the indexes are much less than 0.05 which suggests that all the points are outliers and needs to be investigated and removed if required to make the model perform better.

4.7 Backward Regression

The backward step-regression has been performed to check whether the best fit model requires all the predictor variables or any of the variables can be removed. The results have been shown below.

```
Start:  AIC=154312
price1 ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors +
    waterfront + view + grade + condition + City

              Df Sum of Sq       RSS    AIC
<none>                     434174484 154312
- sqft_lot     1    433120 434607604 154325
- bedrooms    10   2908752 437083236 154392
- floors       5   5341378 439515862 154485
- condition    4   9241858 443416342 154620
- view         4  11770935 445945419 154705
- bathrooms   27  26170357 460344841 155136
- waterfront   1  28965087 463139571 155279
- sqft_living  1  50119462 484293946 155949
- grade       10  69197175 503371659 156510
- City        48 232028931 666203415 160638
```

The table clearly shows that removing any of the variables increases the AIC (Akaike Information Criteria) i.e. makes our model weak. Therefore, we are not removing any of the response variables and conclude that this is the best fit model.

4.8 Prediction on Test Data

The entire data was split into 70:30 ratio and the model were built on the 70 percent data. The remaining 30 percent data has been used to check how the model predicts the house prices and the corresponding prediction variation with respect to actual price.

The accuracy has been calculated based on two criteria:

$$MinMaxAccuracy = mean\left(\frac{min\,(actuals, predicteds)}{max\,(actuals, predicteds)}\right)$$

$$MeanAbsolutePercentageError\,(MAPE) = mean\left(\frac{abs\,(predicteds - actuals)}{actuals}\right)$$

The MinMaxAccuracy is found to be 0.82 which tells that the model is able to predict the house prices almost correctly 82 percent of the times. The MAPE is found to be 0.21 which suggests that there is a prediction deviation of 21 percent i.e. there is quite a few prices predicted inaccurately.

5. Conclusion and Future Work

In this paper, we tried to build a linear regression model to predict the house prices of King County. We created a best fit model that can predict the house prices with reasonable accuracy. Although the results show that the model is able to predict the price of the houses with good accuracy but there are plenty of aspects that can be improved to make it perform better. The exploratory analysis on the data suggests that it is highly skewed and contains many outliers. Therefore, to improve the model, skewness of the data needs to be rectified and the outliers must be investigated and removed. Moreover, the data splitting needs to be done in such a way that all the conditions are covered in the training dataset. In future we can make all these necessary modifications that will improve the model accuracy. Further diagnostics needs be done on the model for performance improvement which will be explored in the future.

6. References

[1] Guangliang Gao, Zhifeng Bao, Jie Cao, A. K. Qin, Timos Sellis, Fellow, IEEE and Zhiang Wu, “Location-Centered House Price Prediction: A Multi-Task Learning Approach”.

[2] R. YAYAR and D. DEMIR, “Hedonic estimation of housing market prices in turkey,” Erciyes ¨Universitesi Iktisadi ve Idari Bilimler Fak¨ultesi Dergisi, no. 43, pp. 67–82, 2014.

[3] S. Rosen, “Hedonic prices and implicit markets: product differentiation in pure competition,” Journal of political economy, vol. 82, no. 1, pp. 34–55, 1974.

[4] Mongare G. Kemunto and. Dr. Wilfred Nyangena, “RESIDENTIAL HOUSING DEMAND IN NAIROBI; A HEDONIC PRICING APPROACH”.

[5] M. Kryvobokov and M. Wilhelmsson, “Analysing location attributes with a hedonic model for apartment prices in donetsk, ukraine,” International Journal of Strategic Property Management, vol. 11, no. 3, pp. 157–178, 2007.

[6] https://www.kaggle.com/swathiachath/kchousesales-data